

AD_____

Award Number: W81XWH-13-C-0042

TITLE: Learning the Language of Healthcare Enabling Semantic Web Technology in CHCS

PRINCIPAL INVESTIGATOR: Dr. Wesley Turner

CONTRACTING ORGANIZATION: Kitware, Inc.
Clifton Park, NY 12065

REPORT DATE: September 2013

TYPE OF REPORT: Final @

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release; Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE September 2013	2. REPORT TYPE Final Phase I	3. DATES COVERED 18 February 2013-18 August 2013		
4. TITLE AND SUBTITLE Learning the Language of Healthcare Enabling Semantic Web Technology in CHCS		5a. CONTRACT NUMBER W81XWH-13-C-0042		
		5b. GRANT NUMBER		
		5c. PROGRAM ELEMENT NUMBER		
6. AUTHOR(S) Dr. Wesley Turner		5d. PROJECT NUMBER		
		5e. TASK NUMBER		
		5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Kitware, Inc. Clifton Park, NY 12065		8. PERFORMING ORGANIZATION REPORT NUMBER		
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012		10. SPONSOR/MONITOR'S ACRONYM(S)		
		11. SPONSOR/MONITOR'S REPORT NUMBER(S)		
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited				
13. SUPPLEMENTARY NOTES				
14. ABSTRACT This document represents the Final Report of the SBIR Phase I effort under W81XWH-13-C-0042. During the Phase I effort we successfully demonstrated the capability to extract semantic information from File Man based Electronic Health Record implementations and to use the extracted information in informative visualizations and analyses. Additional efforts on the Phase I introduced our concepts to our potential collaborative community, investigated and categorized the semantic health care ecosystem, and planned for the Phase II implementation of a prototype communications capability				
15. SUBJECT TERMS- SBIR Report, Learning the Language of Healthcare Enabling Semantic Web Technology in CHCS, Unclassified				
16. SECURITY CLASSIFICATION OF: a. REPORT U		17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 13	19a. NAME OF RESPONSIBLE PERSON USAMRMC 19b. TELEPHONE NUMBER (include area code)
b. ABSTRACT U		c. THIS PAGE U		

Phase I SBIR Final Report:

Learning the Language of Healthcare

Enabling Semantic Web Technology in CHCS

Table of Contents

Results of the Phase I Work	4
Technical Objectives for the Phase I Effort	4
Open Source Development Environment and User Community - Aim 1	4
Open Source Development Tools	5
Open Test Platform	5
Outreach Activities	6
Integration of Semantic Web Components - Aim 2	6
Database Importation	6
FMQL Introspection Layer	7
Visualization of FMQL results	10
Current Environment and Gap Analysis - Aim 3	11
Conclusions	11
References	13

Results of the Phase I Work

Technical Objectives for the Phase I Effort

During this SBIR, we will develop and demonstrate a working, open source, and patient-centric language using Semantic Web technologies and employing the Resource Description Framework (RDF), Web Ontology Language (OWL), and query technologies such as SPARQL, which utilize the concept of “tuples”, (subject, predicate, object), to relate data and achieve semantic interoperability. Other similar technologies exist, but their proprietary nature prevents their wide adoption. In contrast, our open source ecosystem will provide a foundation for cross-institutional information exchange and will facilitate data mining for medical research purposes and encourage wide adoption.

For Phase I, we defined the following three Specific Aims:

1. Create a sandbox with the components of an open-source development project and seed the community.
2. Gather and integrate existing and relevant open source Semantic Web components into the sandbox
3. Determine and explore key areas missing from existing work. At a minimum, these areas include access control, provenance, and how CHCS differs from VistA.

In Aim 1, we proposed to develop a working demonstration “sandbox” that could be used as a framework for the Phase II prototype and as a concrete example of the system concepts. To facilitate the use of the sandbox, we proposed to initiate an open public semantic repository based on our team’s experience with VistA analytics and to build a community of adopters to extend the reach of the platform to a multitude of collaborating institutions. Some of the tools that we proposed include github for revision control, and wikis for community coordination.

In Aim 2, we proposed to focus on the execution of a practical and realistic infrastructure suitable for wide adoption and rapid spread. As we have in previous large-scale projects, we proposed to exercise agile methodologies based on peer-production models to tackle problems of large scale. We also proposed to begin the process of building a user community focused on the use of the software in practical clinical, research, and personal health activities. In the same way that the Wikipedian community has eclipsed the technology of the wiki, we seek to enable a robust community, applying this technology and feeding back suggestions for future use.

In Aim 3, we proposed to evaluate the existing body of research and working systems that we believed are already sufficient for describing healthcare information to determine missing functionality or new development required for growth and wide adoption. We proposed to accomplish this using a series of workshops, social media, and presentation at conferences.

Based on the findings resulting from these three objectives, we intended to define a design to be fully implemented in Phase II. The rest of this section details the steps we have taken to fulfill these Specific Aims and our accomplishments to date.

Open Source Development Environment and User Community - Aim 1



Open Source Development Tools

Early in Phase I, we began the task of introducing Semantic Healthcare as an open source project by creating a project on github [4]. The front page, shown in Figure 1, contains an overview of the project and provides a link [5] to the tools provided including a set of repositories, an issue tracker, a pull request capability, and a wiki page. Github also provides tools that allow the site usage to be tracked and analyzed. Within the site, we currently maintain three repositories for *SemanticHealthcare.github.io*, *SemanticHealthcare*, and *SemanticHealthDemo*

representing the code for the

front page, experimental code for the development activities, and some demonstration visualizations, respectively. Our team actively used the repositories for communication of current code status and to maintain revision history. We believe that this will positively impact our ability to attract additional community members as the project matures and enters Phase II.

At the same time that we set up the github account, we also purchased rights to the domain names *SemanticHealthcare.biz*, *SemanticHealthcare.us*, *SemanticHealthcare.info*, and *SemanticHealthcare.net*. These domains will allow us to brand our system and to reach out to the community in a more effective way than relying solely on the github repository.

Open Test Platform

The development of the Phase I research code required running EHR system software that we could interrogate and introspect. We were unable to get the access we had hoped to a running CHCS system, but we had full access to the OSEHRA VistA EHR code. Since the two code bases share a common root and because our goal is to provide an interconnect capability adaptable to multiple EHRs, we decided that the VistA alternative, as proposed in the Phase I contingency plan, was viable. Subsequently, we set up running instances of the OSEHRA VistA EHR on two machines on *Rackspace*. These two machines were allocated the purchased domain names *SemanticHealthcare.net* and *SemanticHealthcare.info*.

At points during our Phase I activities, these installations were modified to include additional data and to support FMQL queries. Our complete instructions for setting up a comparable system can be found in our repository [6] and should be available for other researchers. Note that we added data several times to the installations of OSEHRA VistA including data from database files provided by our sponsor. Although our current machines are now open to external queries, we did not allow unauthorized access to sponsor data at any point.

Outreach Activities

Outreach is an essential part of trying to grow an open source project as well as to learn about the viewpoints of thought leaders in the field. Despite our limited Phase I development schedule, we carried out a number of outreach activities spanning 4 workshops in San Diego, San Francisco, and MIT.

The workshop “Science of the Individual” (San Diego, Nov 2012) [11] focused on the issues of a patient-centric framework for a Universal Health Language. It explored the role of genomics, patient privacy, and the dynamics of going through a period of accelerating change in healthcare, but with a healthcare IT infrastructure that seems to be decelerating in its ability to change. We discussed the necessity of a health meta language to cope with these complexities. This was followed by the “Semantic Health” workshop at the headquarters of the World Wide Web Consortium at MIT Apr 19-20, 2013 [12] where the notion of RDF as a candidate for the “Universal Health Language” was crystallized. At the “Semantic Healthcare: workshop, we also discussed how the fine granularity and simple regularity of the RDF approach could enhance patient privacy and security and support the necessary provenance metadata.

Our next outreach activity was a June, 2013 workshop entitled “RDF as Universal Health Exchange Language” at the Semantic Technology Conference, San Francisco [13]. One of the outputs from the workshop was Tom Munnecke’s Interview with pioneering Stanford Health IT semantic web expert Gio Widerhold [14]. This interview generated considerable attention on various mailing lists, and is now slated to be part of the National Library of Medicine’ oral history of Health IT.

The final workshop for this effort was held in June 2013: “RDF as Universal Health Language” in San Diego [15] which brought in theoreticians reexamining Health information from a Category Theory perspective, as well as practitioners, including the CMIO of Kaiser Permanente, talking about practical experiences with medical nomenclature and the VA/Kaiser NHIN interface.

Note that while the June conference represented the last of the workshops, we did have one more notable outreach activity. In September 2013, Tom Munnecke served as the coordinator of a panel at the Second Annual OSEHRA Summit in Bethesda, MD entitled “Semantic Approaches to Interoperability”. On the panel, Tom, Conor Dowling, and Dr. Wesley Turner all presented views and results of this Phase I activity.

Beyond the specific topics of these workshops, our outreach activities are also creating a diverse network of thinkers across a broad range of disciplines. Some are academicians, some are bloggers, some are senior level managers and some are scientists. This has provided a strong foundation for discussing and understanding the issues of Phase I, but it will also provide continuity for stage II.

Integration of Semantic Web Components - Aim 2

Database Importation

There are several aspects to the analysis of an EHR installation. Some basic information on the schema can be extracted based solely on the relationships in the EHR files. These relationships and pointers exist regardless of the existence of patient data. However, this type of static analysis cannot give a full picture of how the patient data is actually used, nor can it demonstrate the capability we need to extract and serialize patient data for communication to other EHRs. In order to more fully test and demonstrate data extraction, we needed to find a source of patient data. We took two different approaches to filling this requirement.

The first approach was to use a synthesized set of data that could be automatically loaded into an empty EHR. This approach is identical in intent with existing OSEHRA utilities used to populate OSEHRA VistA instances prior to and during OSEHRA testing operations. Our approach extended this capability by coupling the OSEHRA utilities to a set of fictional patient data encoded in CSV files using Python scripts; and by expanding the categories of data that could be encoded. Our scripts have been placed in the Semantic Healthcare Github repository [1]. The data files are unpublished as they are the property of the sponsoring organization and cannot be used outside of the SBIR activities; however, this general approach can be expanded upon to give an arbitrarily complex set of interrelated patient data if desired.

The second approach was to utilize existing patient demo data that are available in the public domain. The CPRS Demo database has been utilized by the VA for years to showcase their VistA EHR software, it contains thousands of fictionalized patient data entries with abundant encounters, problem lists, diagnosis, vitals, etc., and it is publicly available via VA's FOIA download site [7]. However, this VA database only runs on Intersystems Caché, a proprietary M[umps] platform. As part of the SBIR effort, we ported the CPRS demo data to our GT.M/Linux based platform and set up an RPC broker under the xinetd daemon. Detailed documentation on how we accomplished this port can be found in the Semantic Healthcare repository [5].

Ultimately, both of our data approaches were successful. However, our current test system is based on the CPRS demo data for two primary reasons. First, the CPRS demo had a large, immediately available, set of patient data comprising:

- 1505 Users
- 1630 Patients
- 648 Labs
- 3429 Prescriptions.
- 848 Allergies
- 58 Immunizations
- 938 Problems
- 32467 Vitals
- 2880 Documents
- 26709 Orders
- 409 Locations

Second, the sponsor provided CSV data cannot be used past the end of the SBIR and could not be made public. This limited our ability to provide our test systems to the open source community for feedback and additional development, and limited our ability to continue our efforts past the end of the SBIR.

FMQL Introspection Layer

VistA has many different mechanisms for extracting and representing subsets of its data including multiple versions of HL7 and over 2500 broad and highly granular RPCs. With the open-source FileMan Query Language (FMQL), all data in a VistA's NoSQL store, FileMan, is accessible through one query mechanism and serialized in one RDF format.

Key to FMQL exposure is that each and every piece of information in a VistA gets a unique portable identifier - a web-resolvable URI which distinguishes it from its peers within that VistA and in other systems. This unique identifier is composed of two parts. At the top is the internet domain name of the EHR host. At the bottom is the unique record identifier within the EHR database. For example, our test FOIA VistA instance has the registered domain name "semantichealthcare.info" and VistA keeps patient vital results in file 120.5. As a result, the first patient vital result gets the URI, "http://semantichealthcare.info/rambler#120_5/1". No other piece of data from this VistA, CHCS, Epic, etc. would have this same id. As a result, vitals from this system can easily be distinguished, merged, translated and presented, unambiguously in one health-care "information space". This type of multi-level representation is consistent with the organization of the internet and with other uniquely identified data including DICOM medical images. During the course of this Phase I, we were able to demonstrate this ability to uniquely identify, extract, and represent VistA data during our explorations of the data using, for example, our Patient Data, Institution, Terminology ("Know-how") and System

data (PIKS) reports [16]. Figure 2 shows one of these patient slice reports.

Our investigations proved our ability to represent and explore EHR data using our mechanisms and assumptions. The FMQL projection makes clear that FileMan holds highly interlinked data. And like all networks of data, FileMan's has anchor nodes around which the data cluster. In VistA, 2500 types of data are drawn around two key "nodes" - the "Patient" file holding patient specific information and the "New Person" file holding the User and Provider information. This natural organization partitions and arranges the data in FileMan. Figure 3 below shows the twenty percent of VistA nodes that gather around Patient and demonstrates that most such nodes refer directly to the patient data. The data is tightly clustered. Only one file is four degrees of separation away.

Patient Slice

Those files that hold Patient Data - other slices hold Institution, Terminology ("Know-how") and System data (PIKS).

For native FileMan 'a file is a file' ("Ambiguous File Type problem") - there is no difference between a log file and patient data, but for a graph representation of FileMan data, such distinctions are key.

For graphing, another problem ("Ambiguous Multiple problem") is native FileMan's lack of distinction between multiples (contained files) with nuanced data ('top multiples'), multiples that form lists ('list multiples') and multiples that group fields ('blank nodes').

In the Patient Slice we surface a graph-friendly, Patient Data Model for a FileMan system.

In this system of 2501 top files, 475 hold patient data, of which 27 are populated. 131 top multiples are also in the Patient slice.

Note that no system uses all the files in its FileMan. Empty files - those with no count in the tables below - are placed at the end of each table and can be ignored when querying this particular system. There are five tables below, showing different types of Patient file in order of importance.

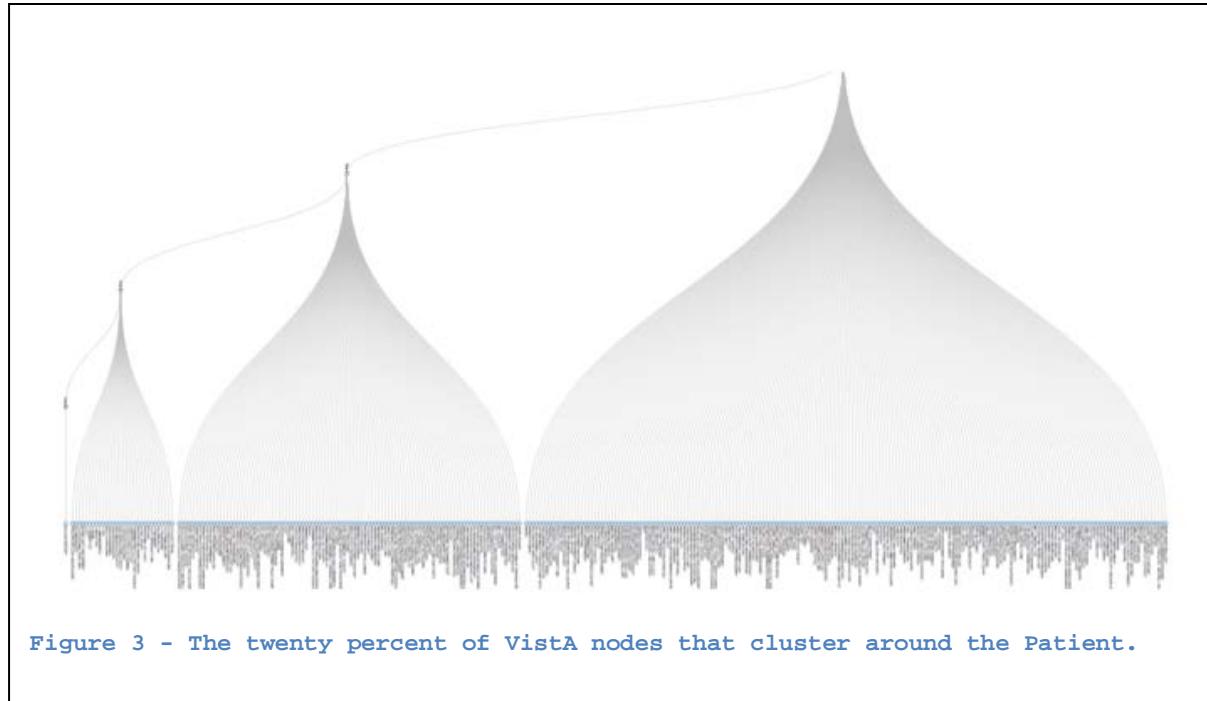
Note - explicitly excluded file(s): 60, 200, 9_2, 4, 62, 101, 95_3, 363_2, 61, 44, 68, 64, 441, 123_5, 62_06, 799_6, 695_5, 810_4, 6695_9, 133, 3_9, 62_3, 15_2, 58_81, 200. FileMan has some 'rogue' / misleading' references but most invalid entries come from including containers of patient data rather than the key multiples they hold.

Direct and Indexed

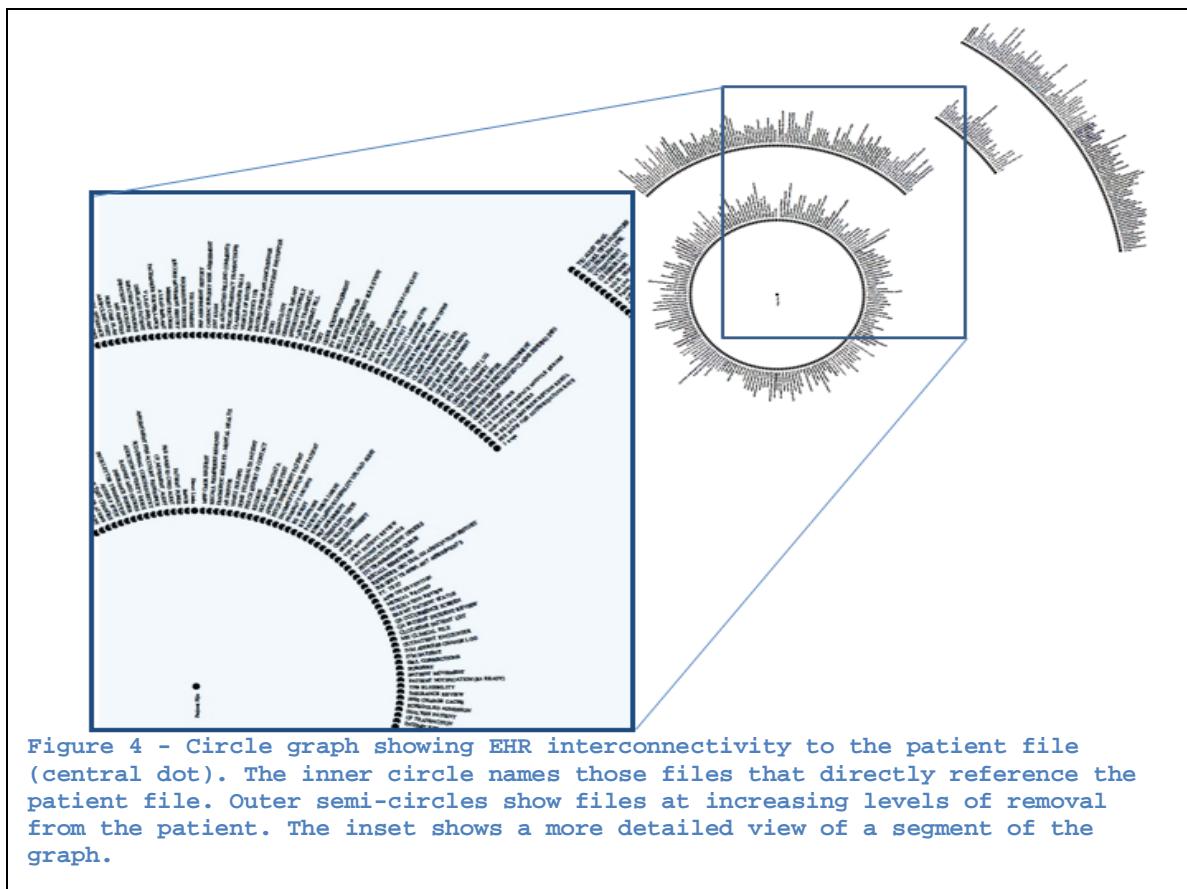
Files with direct, indexed references to the Patient file. This covers most key patient files.

	File	# in System	# Fields
1	Alert Tracking (8992_1)	40	16
2	Patient/Ins (90000001)	27	3
3	Ivm Financial Query Log (301_62)	12	11
4	Ivm Patient (301_5)	11	18
5	Ror Patient Events (798_3)	6	3
6	Outpatient Encounter (409_68)	5	25
7	Patient Enrollment (27_11)	4	38
8	Sd Wait List (409_3)	1	50
9	Patient Allergies (120_8)	1	22
10	Ivm Address Change Log (201_7)	1	21

Figure 2 - A generated patient slice from our FMQL interface. Records within the report are uniquely linked to underlying patient data and EHR records using UIDs.



But clustering around key nodes is only the first, most obvious way VistA data arranges itself. Nodes



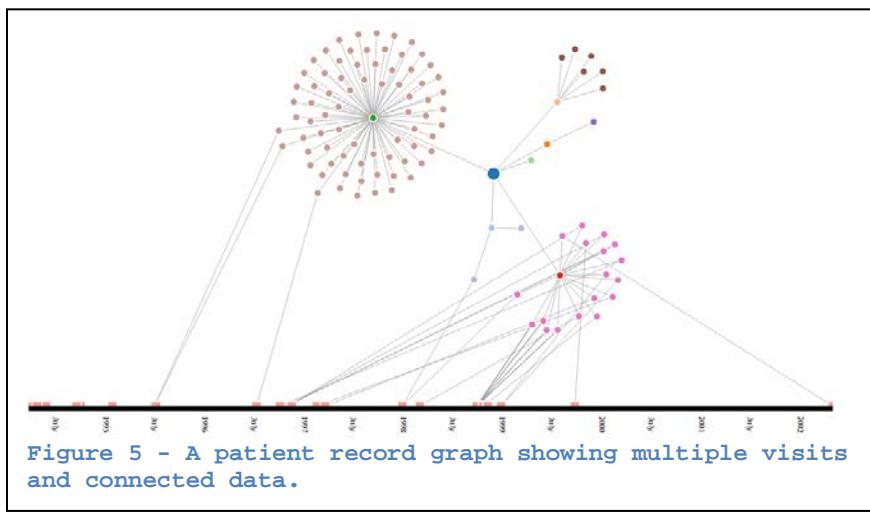
linked, directly or indirectly to Patient represent only a fifth of VistA file types. The remaining nodes record know-how (medications orderable, lab tests, allergens), system configurations, logs, and the setup of an institution itself - the wards, radiology rooms and providers that must be described before a patient enters a hospital. And within Patient linked data, there's longitudinal data - the vitals, prescriptions and problems of a patient at a given point in time - and the workflow that leads to it - doctor's order and consultation requests.

Our investigations confirm our ability to extract this information from an EHR and to represent it in a serialized form with unique identifiers for further processing.

Visualization of FMQL results

Kitware has a long history of support for scientific visualization and a growing presence in the informatics visualization field. We often find that the extraction of information from data is greatly enhanced by asking the right question and finding the right view. Consider Figure 4. This visualization of a tree layout graph [3] shows the relationship between the patient file in VistA and the other data files organized by distance from the patient file. The inner ring shows files that directly reference the patient file; the second ring out shows files reachable through one intermediary; the third ring shows files referenced through two intermediaries; and so on. This type of visualization is useful for analyzing the static relationships among the possible stored data. It tells us nothing of the actual data stored for a patient, but does tell us usable information about the schema and data organization.

Now consider Figure 5. This force directed graph [2] display the most pertinent information about a specified patient from a VistA instance. When a patient is selected, a series of FMQL queries are run to retrieve the information as JSON data objects which are processed and used as the node data of the Force Directed graph. Currently, seven types of information about each patient are evaluated: Visits, Vitals, Allergies, Immunizations, Problems, CPT Codes, and Prescriptions. The largest circle in the middle of the visualization represents the patient, the nodes that have direct links to the patient



node represent data labels, and nodes connected to the labels are the actual patient information. A timeline along the bottom of the visualization represents the gap of time between the first and last visits of the patient and all visit nodes are placed in the correct location along the timeline. Any data item sharing a date with a visit is connected to the

visit by a link.

While this example is only an initial conception, the graph demonstrates a few interesting characteristics. The large “puff ball” to upper left of the graph contains patient vitals. The general lack of connection between vitals and visits either indicates shortcomings of the synthetic data, or

represents an unexpected workflow that needs to be explored. As we move into additional research, these types of questions will become important in the semantic interpretation of the data.

Current Environment and Gap Analysis - Aim 3

As a final component of our research, we undertook a comprehensive literature review of current technologies in the semantic healthcare environment. The document, "Semantic Healthcare", is written in LaTex. A PDF of the document can be found at [9], while the source is archived in our open source repository [10]. The document is a wide ranging analysis of available technologies and current initiatives and provides a guide for us as we consider paths for continued development.

Several technologies are particularly relevant for our continued effort. The Resource Description Framework (RDF) is a World Wide Web Consortium (W3C) standardized data model for representing semantic web resources. RDF provides a graph based self-describing semantic approach supporting automated inference that makes it ideal as a healthcare exchange language. All entities in this model can be uniquely identified by Internationalized Resource Identifiers (IRIs) and include links to other IRIs that enable linking of different resources. Linking data to other web resources and terminologies can be achieved using popular machine readable ontologies such as Friend of a Friend (FOAF), Dublin Core Metadata Initiative (DCMI) and PROV-O. W3C Vocabulary of Interlinked Datasets (VOID) describes these linkages and forms a bridge between the providers and users of RDF data.

Transmission and interchange are also important considerations. Health Level 7 (HL7) and Nationwide Health Information Network (NwHIN) provide a framework for the secure exchange, integration, sharing, and retrieval of electronic health information. These standards define how information is packaged and communicated from one party to another, setting the language, structure and data types required for seamless integration between systems. HL7 Fast Healthcare Interoperability Resources (HL7 FHIR) are a set of atomic and independent resources that represent clinical concepts for exchange to quickly and effectively solve problems in healthcare process and communications. NwHIN provides methods to perform universal patient lookup, document discovery and retrieval and exchange between organizations and federal agencies including VA and DoD. NwHIN CONNECT is an open source health information exchange platform that facilitates secure interoperable health information exchange among diverse technologies.

Having discussed the RDF data model, the HL7 FHIR standards and resources and NwHIN CONNECT secure exchange framework, the last piece in the architecture is a medical terminology lexicon. 3M™ Healthcare Data Dictionary (HDD) is a controlled medical vocabulary server that makes it possible to map and manage different medical terminologies, integrate content and standardize healthcare data. 3M HDD can be accessed via HL7 Common Terminology Services (CTS) that allows clinical databases to exchange, compare, query and connect meaningful data.

While the above discussed standards and technologies prove independently useful towards a unified EHR system, additional efforts are required to implement an integrated application.

Conclusions

We believe that this Phase I SBIR has been successful in achieving our three Specific Aims. As described in the preceding paragraphs, we have set up the conditions for the adoption and growth of an open source community using our websites, open source development tools, test fixtures and outreach activities. These steps have defined a critical core for participation and have begun to seed the community with information on our approach and goals.

These open source components have also been used as a central repository for dissemination of proof of concept tools and community facing test fixtures to include running EHR instances populated with synthetic and openly available test data. We were able to use our tools to extract system dependencies and workflows; to extract semantically related patient data; and to browse patient-centric views into the system. We were also able to use the extracted data using Kitware web visualization tools to provide user friendly depictions of the contents of the data files themselves. Most notably, our tools were able to demonstrate some of the functional inadequacies and inconsistencies in the synthetic data giving us additional insight and providing a basis for improved data sets during our additional investigations.

Finally, our outreach and our continued research allowed us to author two documents, an “Architectural Vision” [8] document and a “Semantic Healthcare” [9] status report. These two documents lay out our vision of the next phase of the project and provide a summary of available tools and applications that we can leverage in our future development. As we expected, it appears that most of the tools and specifications that we will need to implement a demonstration capability currently exist in the Semantic Health ecosystem.

References

- [1] <https://github.com/SemanticHealthcare/SemanticHealthcare/tree/master/ImportScripts>
- [2] <https://github.com/mbostock/d3/wiki/Force-Layout>
- [3] <https://github.com/mbostock/d3/wiki/Tree-Layout>
- [4] <http://semantichealthcare.github.io/>
- [5] <https://github.com/SemanticHealthcare/SemanticHealthcare>
- [6] <https://github.com/SemanticHealthcare/SemanticHealthcare/blob/master/SetupCPRSDemo.rst>
- [7] https://downloads.va.gov/files/FOIA/Software/DBA_VistA_FOIA_System_Files/cprsdemoCACHE.zip
- [8] <http://semantichealthcare.github.io/A%20Vision%20of%20a%20Semantic%20Health%20Care%20Information%20Architecture.pdf>
- [9] <http://semantichealthcare.github.io/SemanticHealthcare.pdf>
- [10] <https://github.com/SemanticHealthcare/SemanticHealthcare/tree/master/ResearchReport>
- [11] <http://www.new-health-project.net/2013/03/21/workshop-report-science-of-individual/>
- [12] <http://www.new-health-project.net/2013/04/23/report-on-semantic-health-workshop-at-mit-april-19-20-2013/>
- [13] <http://www.youtube.com/playlist?list=PLt-A972QBADUZZK9tlzNhDSPb6gv-Uvog>
- [14] <http://www.youtube.com/watch?v=rpC9dDYrquw>
- [15] <http://www.youtube.com/playlist?list=PLt-A972QBADVxzo9hgz87R9QVrWUkrb>
- [16] <http://www.semantichealthcare.net/reports/PIKS/patientSlice.html>